ARTICLE **Communicated by Ron Meir**

Operant Matching as a Nash Equilibrium of an Intertemporal Game

Yonatan Loewenstein
yonatan@huji.ac.il
Departments of Neurobiology and Cognitive Sciences and the Interdisciplinary Center for Neural Computation, Hebrew University, Jerusalem, 91904, Israel

Drazen Prelec
dprelec@mit.edu
Sloan School of Management, Massachusetts Institute of Technology, Cambridge MA 02139, U.S.A.

H. Sebastian Seung
seung@mit.edu
Howard Hughes Medical Institute and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.

Over the past several decades, economists, psychologists, and neuroscientists have conducted experiments in which a subject, human or animal, repeatedly chooses between alternative actions and is rewarded based on choice history. While individual choices are unpredictable, aggregate behavior typically follows Herrnstein's matching law: the average reward per choice is equal for all chosen alternatives. In general, matching behavior does not maximize the overall reward delivered to the subject, and therefore matching appears inconsistent with the principle of utility maximization. Here we show that matching can be made consistent with maximization by regarding the choices of a single subject as being made by a sequence of multiple selves—one for each instant of time. If each self is blind to the state of the world and discounts future rewards completely, then the resulting game has at least one Nash equilibrium that satisfies both Herrnstein's matching law and the unpredictability of individual choices. This equilibrium is, in general, Pareto suboptimal, and can be understood as a mutual defection of the multiple selves in an intertemporal prisoner's dilemma.

The mathematical assumptions about the multiple selves should not be interpreted literally as psychological assumptions. Human and animals do remember past choices and care about future rewards. However, they may be unable to comprehend or take into account the relationship between past and future. This can be made more explicit when a mechanism that converges on the equilibrium, such as reinforcement learning, is considered.

Neural Computation **21, 2755–2773** (2009)

Using specific examples, we show that there exist behaviors that satisfy the matching law but are not Nash equilibria. We expect that these behaviors will not be observed experimentally in animals and humans. If this is the case, the Nash equilibrium formulation can be regarded as a refinement of Herrnstein's matching law.

1 Introduction

There is a long tradition of laboratory experiments in which animal and human subjects make repeated decisions that can result in reward. Although individual choices in these experiments are typically unpredictable, aggregate behavior in a variety of reward schedules is often well approximated by the matching law. This empirical regularity states that over a long series of repeated trials, the reward accumulated from choosing an action is proportional to the number of times the action was chosen (Rachlin & Laibson, 1997; Gallistel, Mark, King, & Latham, 2001; Sugrue, Corrado, & Newsome, 2004; Davison & McCarthy, 1988). For a precise statement of the matching law, suppose that the experiment consists of a large number T of discrete trials in which the subject chooses among N possible actions.[1] We denote the reward received in trial t by $R(t)$ and define the N-dimensional binary vector $\mathbf{A}(t)$ by

$$A_i(t) = \begin{cases} 1, & i\text{th action is chosen at trial } t \\ 0, & \text{otherwise.} \end{cases}$$

Definition 1: The matching law. *There exists a constant $\bar{R}$ such that*

$$\frac{1}{T}\sum_{t=1}^{T} A_i(t)R(t) = \bar{R}\frac{1}{T}\sum_{t=1}^{T} A_i(t) \tag{1.1}$$

for all actions i.

The left-hand side is the time-averaged reward received in trials in which action i was chosen. The right-hand side is proportional to the fraction of times action i was chosen. It is easy to see that $\bar{R}$ is the time-averaged reward per trial.[2]

[1] The matching law was originally investigated for reward schedules in which actions and rewards occur in continuous time. For simplicity we consider the case of discrete time, which was introduced in the 1980s (Herrnstein, Prelec, & Vaughan, 1986).

[2] In practice, it may take some time for the behavior to equilibrate, so the initial trials of the experiment are sometimes left out of the summation.

The emergence of matching behavior is typically explained algorithmically using dynamical models. In these models, the subject makes choices stochastically, as if by tossing a biased coin. The choice probabilities evolve in time based on the actions chosen and rewards received such that the choice probabilities converge to values satisfying the matching law (Sugrue et al., 2004; Gallistel, Mark, King, & Latham, 2002; Gibbon, 1995). For example, according to the theory of melioration (Herrnstein & Prelec, 1991), subjects estimate the return from the different alternatives and shift their choice preference in the direction of the alternatives that provide a higher-than-average return.[3] In a diminishing return reward schedule, the return from an alternative decreases with the frequency of choosing that alternative in the past. Therefore, the shift in choice preference in favor of an alternative reduces the return from that alternative. This dynamical learning process reaches a fixed point when choices are allocated such that the return from all chosen alternatives is equal. While the theory of melioration provides an important insight into how matching behavior may emerge, the computational principles underlying this learning algorithm are not clear, in particular because in general, matching behavior does not maximize the overall reward delivered to the subject. In fact, the relationship between the matching law and the behavior that maximizes the overall reward has been a subject of debate for several decades (Herrnstein & Loveland, 2001; Heyman & Luce, 1979; Vaughan, 1981; Mazur, 1981; Vyse & Belke, 1992; Sakai & Fukai, 2008).

Normative economic theories that discuss choice behavior typically view subjects as deliberative agents that maximize a utility function under the constraints that they face. We seek to find whether there is any sense in which stochastic matching behavior can be justified in this framework, independent of a learning algorithm. More precisely, we seek to find a set of assumptions on the information used by the subject when choosing his actions and on his objective function such that the empirically observed matching law maximizes this objective function given this available information. For this purpose, we use the Markov decision process (MDP) formalism to model the rewards received by an agent interacting with a dynamic world. In order to solve for the maximizing behavior, the MDP is regarded as a game between multiple selves, each choosing an action and receiving the reward at a single time. With the assumptions that the agent has no knowledge of the state of the world and discounts the future completely, we prove that there exists a Nash equilibrium of the game in which every action is chosen independently at random from the same probability distribution. Furthermore, this Nash equilibrium satisfies the matching law,

[3] Gallistel et al. (2001) have criticized this general approach, arguing that it is inconsistent with rapid changes in behavior following changes in reward schedules. Instead, they favor a model based on estimation of time intervals (Gallistel et al., 2002). (However, see also Neiman & Loewenstein, 2007.)

and this is a sense in which matching can be "justified." This formulation is illustrated by computing the Nash equilibrium for several example MDPs that have been used in the laboratory as reward schedules or have been previously proposed as models of hedonic psychology.

The mathematical assumptions about the multiple selves should not be interpreted literally as psychological assumptions. Humans and animals do remember past choices and care about future rewards. However, they may be unable to comprehend or take into account the relationship between past and future. In section 9, we examine possible reasons why subjects behave as though they make these assumptions.

The multiple-selves approach has apparently paradoxical implications. For a generic MDP, the selves could receive higher expected payoffs if they cooperated to choose from some other probability distribution, but this strategy would not be stable to defection by a single self. In contrast, the matching strategy is stable to defection but typically ends up being Pareto suboptimal. This shows how matching can seem both rational and irrational, similar to mutual defection in the prisoner's dilemma. Our article formalizes the previous suggestion by Herrnstein et al. (1986) that matching is similar to an intertemporal prisoner's dilemma.

The Nash equilibrium formulation of the matching law bears two important implications. First, the Nash equilibria of the multiple-selves game are only a subset of the solutions of the empirical matching law (see equation 1.1). Typically, solutions to equation 1.1. that are not Nash equilibria are not observed experimentally in human and animal subjects. Thus, our formulation can be regarded as a refinement of the matching law. Second, our formalism contributes to the decades-long debate on the relationship between matching and maximizing by explicitly pointing out a set of assumptions on the information used by the subject and on his objective function such that the empirically observed matching law maximizes the objective function given the available information.

Some of the findings presented here have appeared previously in abstract form (Loewenstein, Prelec, & Seung, 2007).

2 Optimal Control

We will rely on the Markov decision process (MDP) formalism to model the relationship between actions and rewards. This formalism is extremely general and includes all the reward schedules that have been used in the laboratory to study matching.

Definition 2. *An MDP is defined by a transition function and a reward function over state and action spaces. The transition function* $f(\mathbf{S}(t), \mathbf{A}(t), \mathbf{S}(t+1))$ *is the probability of moving from state* $\mathbf{S}(t)$ *to state* $\mathbf{S}(t+1)$ *after action* $\mathbf{A}(t)$ *is chosen. The reward function* $r(\mathbf{S}(t), \mathbf{A}(t))$ *determines the probability and magnitude of the reward* $R(t)$ *received after taking action* $\mathbf{A}(t)$ *while in state* $\mathbf{S}(t)$.

Note that the state $\mathbf{S}(t)$ depends on the history of past actions. This allows the MDP to have a "memory" of the past, so that actions can have delayed consequences for rewards. (Examples of MDPs are presented in section 6.)

Consider a rational agent that knows the transition function f and reward function r of the MDP. How should the agent behave? The answer to this question depends on the *objective function* of the agent and the observability of the states. Typically it is assumed that the goal is to maximize the expectation value of the sum of present and discounted future rewards. The question of whether a particular policy is "optimal" is crucially dependent on how future rewards are discounted relative to present rewards. It is common to make the discount factor an exponential function of time, and in this case, the optimal policy is a solution of the Bellman optimality equation (Sutton & Barto, 1998). However, hyperbolic discounting may be a better model for animal and human behavior (Frederick, Loewenstein, & ODonoghue, 2002).

The second determinant of optimal policy is the observability of the state. In the simplest case of full observability, the agent knows the state variable $\mathbf{S}(t)$ and can use that information to choose its action $\mathbf{A}(t)$. In a partially observable MDP (POMDP), the agent has access to some information about the state but does not know it completely (McAllester & Singh, 1999). Characterizing optimal behavior in the POMDP case, a problem that is closely related to "decision problem with imperfect recall" in economics, is far more demanding than the fully observable MDP case (Piccione & Rubinstein, 1997). It has been argued that in this family of problems, the optimal plan at the present moment may be suboptimal in the future and thus will not be obeyed (Piccione & Rubinstein, 1997, but see also Aumann, Hart, & Perry, 1997). Therefore, it becomes problematic to speak of the optimal behavior of a single agent. Rather, one should consider a game between multiple selves, one at each time (Ainslie, 1992; Gilboa, 1997).[4] It is generally accepted that optimal behavior in this case is a Nash equilibrium of the game between the selves (Hendon, Jacobsen, & Sloth, 1996; Piccione & Rubinstein, 1997; Gilboa, 1997; Aumann et al., 1997; Hammond, 1997).

Here we consider the case of a nonobservable MDP (NOMDP) in which the agent has no access to state information at all. This implies that the agent has no memory of its past actions, as this memory could be used to infer information about the current state. Since the NOMDP is just an extreme example of a POMDP, the multiple-selves Nash equilibrium will be used as a definition of optimal behavior. We prove that if the temporal discounting in the objective function of the agent is complete, then the Nash equilibrium of the game is consistent with the experimentally observed matching law.

[4]The multiple-selves approach is also the standard solution to the problem of optimal choice in the MDP case when discounting is nonexponential (Laibson, 1997; Strotz, 1955).

3 The Multiple Selves Game

Definition 3. *An NOMDP with infinite discounting can be used to define a multiple-selves normal form game with an infinite number of players—one for each integer time t, $-\infty < t < \infty$. The player t ("self" at time t) chooses action $\boldsymbol{A}(t)$ and receives payoff $r(\boldsymbol{A}(t), \boldsymbol{S}(t))$.*

Because of nonobservability, the sequential character of the MDP is lost. It is as though all of the players choose their actions simultaneously. Because of infinite discounting, the payoff to player t is the immediate reward $r(\boldsymbol{A}(t), \boldsymbol{S}(t))$.

Proposition 1. *A multiple-selves game based on an NOMDP has a time-invariant Nash equilibrium, that is, an equilibrium in which every player draws its action from the same probability distribution $\boldsymbol{p}$: $Pr[A_i(t) = 1] = p_i$.*

This proposition is a special case of theorem 2 in the appendix, which states that any time-invariant game has a time-invariant Nash equilibrium. The NOMDP game is time invariant, since the transition function and reward functions have no explicit dependence on time.

Intuitively, because players do not know what happened before and because all players have the same payoff function, it is reasonable that they all implement the same reasoning and employ the same strategy. In the next section, we show that this strategy is consistent with the experimentally observed matching law.

4 Matching as a Nash Equilibrium

At a time-invariant Nash equilibrium, all actions of the NOMDP are chosen from the same probability distribution $\mathbf{p}$, which defines a Markov process. If the Markov process is ergodic, the average over the stationary distribution of the Markov process is equal to the time averages in the matching law, equation 1.1. Thus, for any action i with nonzero probability $p_i > 0$,

$$\mathbf{E}[R(t)|A_i(t) = 1] = \lim_{T\to\infty} \frac{\sum_{t=1}^{T} R(t)A_i(t)}{\sum_{t=1}^{T} A_i(t)}, \tag{4.1}$$

where the expectation is over the stationary distribution.

At the Nash equilibrium, the expected payoffs of all actions with nonzero probability are equal to the same value. For the multiple-selves game, this means that $\mathbf{E}[R(t)|A_i(t) = 1]$ has the same value $\bar{R}$ for all chosen actions i. Therefore, equation 4.1 implies that the matching law, equation 1.1 is true for all actions i such that $p_i > 0$. Note that the empirical matching law

equation, equation 1.1, holds trivially when $p_i = 0$, because both sides of the equation vanish. This leads us to the following theorem:

Theorem 1. *If the Markov process defined by a time-invariant Nash equilibrium of the multiple-selves NOMDP game is ergodic, the matching law holds.*[5]

It is important to note that while theorem 1 guarantees that under very general conditions, the time-invariant Nash equilibrium of the multiple-selves game is consistent with the phenomenological matching law, equation 1.1, not all solutions of equation 1.1 are Nash equilibria of the game. This is demonstrated in section 6, using two specific examples of a reward schedule.

5 Analogy with the Prisoner's Dilemma

Consider the set of all time-invariant mixed strategies for the multiple-selves game—strategies of the form $\mathbf{p}(t) = \mathbf{p}$. The expected payoff of each player is independent of time, $u(\mathbf{p}) = \mathbf{E}[R(t)]$. Define $\mathbf{p}_{\max}$ as the strategy that maximizes this function. Let's call this a maximizing strategy, since it is reward maximizing in the set of time-invariant mixed strategies. In general, the maximizing strategy differs from the Nash equilibrium, the matching strategy, and has a higher expected payoff, $u(\mathbf{p}_{\max}) > u(\mathbf{p}_{match})$ (see the example in section 6). Thus, the expected payoff of every player under the maximizing strategy is greater than that under the matching strategy, and therefore all players are better off under the maximizing strategy compared to the matching strategy. However, if $\mathbf{p}_{\max}$ is not a Nash equilibrium, then it is unstable to defection by a single player in the sense that if all players play $\mathbf{p}_{\max}$, then a single player can do even better by deviating from $\mathbf{p}_{\max}$. By analogy with the prisoner's dilemma, the matching equilibrium is analogous to the Pareto suboptimal mutual defection, while maximizing is analogous to mutual cooperation. Note that this intertemporal conflict does not result from the nonexponential discounting of future rewards, as infinite discounting can be considered as exponential with an infinite discount rate.

6 Reward Schedules

For more intuition into the Nash equilibrium, it is helpful to consider several specific examples. An experimental prescription for delivering a reward in

[5] In fact, the requirement that the Markov process be ergodic can be significantly weakened. It is sufficient to require that the averages of actions and rewards over the stationary distribution be equal to their average over time, equation 4.1. This equivalence of ensemble average and time average is true for all reward schedules used to study matching behavior experimentally.

response to actions is called a *reward schedule*. In many experiments, the reward, or the probability of reward in a trial, is a function not only of the particular action in that trial but also a function of the W previous actions of the subject,

$$R(t) = r\left(\mathbf{A}(t), \mathbf{A}(t-1), \ldots, \mathbf{A}(t-W)\right). \tag{6.1}$$

The function can be either deterministic, in which case $R(t)$ corresponds to the magnitude of reward, or stochastic, in which case $R(t)$ corresponds to the expectation value of the reward. This is a special case of the MDP in which the state $\mathbf{S}(t)$ is equivalent to the tuple of W previous actions.[6] The rationale behind making reward dependent on past actions is twofold. First, this is a generic property of the real world. For example, the likelihood that a predator will find prey in a particular location depends on its previous actions. It could be smaller if the predator has visited that location recently. Second, the psychology and physiology of the agent may lead to a dependence of subjective reward on past actions. For example, the pleasure of eating an ice cream cone may depend on the time elapsed since ice cream was last consumed. A situation in which choosing an action decreases the future expected reward from that alternative is known in economics as diminishing returns. To be more specific and concrete, suppose that a reward at time t depends on past actions only through their frequencies. Then equation 6.1 specializes to

$$R(t) = r(\mathbf{A}(t), \hat{\mathbf{p}}(t)) = \sum_{i=1}^{N} A_i(t) r_i(\hat{\mathbf{p}}(t)),$$

where the r_i are arbitrary functions and $\hat{p}_i(t) = \frac{1}{W}\sum_{\tau=1}^{W} A_i(t-\tau)$ is the fraction of times action i was chosen in the past W trials. This reward schedule has been used as a reward schedule for behavioral experiments with human subjects (Herrnstein et al., 1986; Herrnstein, Loewenstein, Prelec, & Vaughan, 1993; Egelman, Person, & Montague, 1998; Heyman & Dunn,

[6] Because the experiment has a starting time ($t = 1$), the reward schedule in the first W trials should also be defined. This corresponds to determining the initial probability distribution over the states of the MDP. Note that as a result of the starting time, the corresponding game differs from the multiple-selves game in definition 3 that has no starting point. In the NOMDP case, a self has no information about his temporal location and in principle should take into account the possibility that he is one of the first selves when computing his expected payoff. However, if the number of selves is large, corresponding to the case of a long experiment, the fraction of selves affected by the initial conditions is small. Therefore, the contribution of the initial conditions to the expected payoff vanishes, and the game has the same time-invariant Nash equilibrium as in the case of infinite backward horizon.

2002). In particular, if r_i is a linear function of $\hat{p}_i$ and there are $N = 2$ actions (Herrnstein et al., 1986),

$$R(t) = A_1(t)[a_1 - b_1\hat{p}_1(t)] + A_2(t)[a_2 - b_2\hat{p}_2(t)] \tag{6.2}$$
$$= \begin{cases} a_1 - b_1\hat{p}_1, & \text{if } A_1(t) = 1, \\ a_2 - b_2\hat{p}_2, & \text{if } A_2(t) = 1. \end{cases}$$

The dependence of the return from alternative i, $\mathbf{E}[R(t)|A_i(t) = 1]$ on past actions enters into equation 6.2 through the parameter b_i. A positive b_i indicates a diminishing return reward schedule.

6.1 Two-Armed Bandit. In the two-armed bandit reward schedule, reward in a trial is independent of past actions, $b_1 = b_2 = 0$, and each action provides a binary reward with a fixed probability associated with that action. In general, $a_1 \neq a_2$, and therefore $\mathbf{E}[R(t)|A_1(t) = 1] \neq \mathbf{E}[R(t)|A_2(t) = 1]$. Without loss of generality, we assume that $\mathbf{E}[R(t)|A_1(t) = 1] > \mathbf{E}[R(t)|A_2(t) = 1]$. It is easy to see that in the multiple-selves game, each player will maximize his expected payoff by choosing alternative 1 exclusively, independent of the actions of all other players. Therefore, this game has a single Nash equilibrium, $p_1 = 1$, for all players. In contrast, it is easy to see that there are two policies that are consistent with the empirical matching law, equation 1.1, the Nash equilibrium, $p_1 = 1$, and choosing the inferior action exclusively, $p_2 = 1$. The experimentally observed behavior in sufficiently long experiments with the two-armed bandit reward schedule typically converges to the Nash equilibrium solution, $p_1 = 1$.[7]

6.2 Addiction Model. Consider the reward schedule of equation 6.2 in which $b_1 > 0$ and $b_2 < 0$, depicted in Figure 1. Herrnstein and Prelec (1991) proposed that a similar a reward schedule is a good conceptual model for addiction for the following reason: consider an agent that can choose every day between two actions: "taking drugs" ($A_1 = 1$) or "not taking drugs" ($A_2 = 1$). Addicts typically report that the more often they use drugs, the less pleasure they get out of a single episode of drug use. In other words, there is diminishing return associated with drug use. This property is schematically modeled by the negative slope of the dash-dotted

[7] According to some studies, behavior in the two-armed bandit reward schedule converges to probability matching, a behavior in which the probability of choosing an alternative is proportional to the return from that alternative (Bush & Mosteller, 1955). More recent studies indicate that probability matching is a transient phenomenon, because in longer experiments, behavior slowly converges to the more profitable alternative (Vulkan, 2000; Shanks, Tunney, & McCarthy, 2002). Sustained probability matching may be observed in particular reward schedules that contain a "regret" signal (Gallistel, 1990).

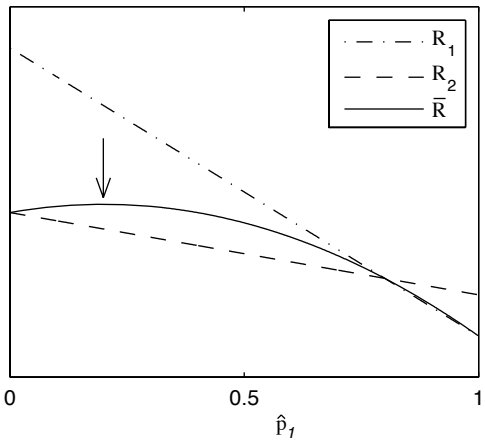

Figure 1: Reward schedule based on choice frequency. The dash-dotted and dashed lines specify the reward obtained from choosing alternatives 1 and 2, respectively, R_1 and R_2, as functions of the frequency $\hat{p}_1$ with which 1 was chosen in the past W trials. Depicted is the case where alternative 1 has a "diminishing return": the reward from choosing an alternative decreases with the frequency of choosing the alternative. The opposite is true for alternative 2. Matching behavior ($\hat{p}_1 = 0.8$) is at the intersection of the dash-dotted and dashed lines. The solid line is the average reward rate, $\overline{R}$, and maximizing behavior corresponds to the peak of the curve ($\hat{p}_1 = 0.2$, arrow). The parameters are $a_1 = 0.9$, $b_1 = 0.7$, $a_2 = 0.3$, and $b_2 = -0.2$.

line in Figure 1, which depicts the "pleasure" from a single "take drugs" (R_1) as a function of frequency of past use, for example, the fraction of days in the past month in which drugs were used ($\hat{p}_1$). Drug addicts also report that the pleasure from other nondrug activities also diminishes with the frequency of drug use (see Herrnstein & Prelec, 1991, for a detailed account of the model). This is modeled by the negative slope of the dashed line, which depicts the "pleasure" from non-drug-associated activities (R_2) as a function of frequency of past drug use ($\hat{p}_1$). Let us find the Nash equilibrium for the model of equation 6.2. Because of the linearity in equation 6.2, the frequencies $\hat{p}_i$ can be replaced by the probabilities p_i. It is easy to see that the time-invariant Nash equilibrium of the multiple-selves game is the mixed strategy p_1 at which $R_1 = R_2$, the intersection of the dash-dotted and dashed lines in the figure ($p_1 = 0.8$). The point $p_1 = 0$, although a solution to equation 1.1, is not a Nash equilibrium, because if all past selves chose $A_2 = 1$ (no drug), then the immediate return from $A_1 = 1$ (drug) is higher than that of $A_2 = 1$ (no drug). Similar argumentation shows that $p_1 = 1$, although a solution to equation 1.1 is also not a Nash equilibrium.

What would be the maximizing stochastic strategy in this model? If the agent makes a decision by tossing a coin with a bias p ($p_1 = p$), then the average reward rate is $\overline{R} \equiv \mathbf{E}[R(t); p_1 = p] = p[a_1 - b_1 p] + (1 - p)[a_2 - b_2(1 - p)]$ (solid line). The maximum of $\overline{R}$, as a function of p, denoted by an arrow ($p_1 = 0.2$), is very different from the Nash equilibrium solution. The difference between the Nash equilibrium and maximizing solutions can be understood in the context of the prisoner's dilemma. Although a situation in which all selves choosing action 1 with a probability $\mathbf{p}_{\max}$ yields a higher return than a situation in which they choose it with the Nash equilibrium probability, it is unstable to defection. If all selves choose $\mathbf{p}_{\max}$, then a single self would be even better off choosing $A_1 = 1$.

7 Melioration and One-State World

The theory of melioration provides an algorithmic account of the matching law. According to this theory, matching is a fixed point of a learning algorithm, in which subjects estimate the return from the different alternatives and shift their choice preference in the direction of the alternatives that provide a higher-than-average return (Herrnstein & Prelec, 1991). To be more concrete, we assume that choices are made by tossing a biased coin such that alternative i is chosen at time t with probability $p_i(t)$, and the change in the probability of choice is proportional to the difference between the return from alternative i and the average return. Formally, $\Delta p_i(t) \propto \mathbf{E}[R(t)|A_i(t) = 1] - \mathbf{E}[R(t)]$. This learning algorithm is a commonly used learning model in economics. Very similar dynamics, known as the replicator equation, is widely used to describe population dynamics in ecology and evolution (Fudenberg & Levine, 1998; Hofbauer & Sigmund, 2003). The neural basis of such learning algorithms has been studied previously (Seung, 2003; Soltani & Wang, 2006; Loewenstein & Seung, 2006; Loewenstein, 2008). Such a learning algorithm can be understood in the framework of the MDP in the following way. Suppose that a subject erroneously believes that the MDP has a single state; that is, he believes that $W = 0$ in equation 6.1. The optimal policy in this n-armed bandit reward schedule is to choose the alternative that provides the highest return exclusively. However, if the returns from the different alternatives are unknown, the subject can use a replicator-like algorithm in order to find this most profitable alternative. Therefore, it may be argued that subjects match because they meliorate, and they meliorate because they erroneously believe that the MDP has only a single state. However, it is important to note that the optimal policy of a subject in a one-state MDP is to choose the better alternative exclusively. Therefore, although the transient learning dynamics of melioration is consistent with the one-state MDP assumption, a steady-state matching behavior, in which $0 < p_i < 1$ is, in general, inconsistent with this one-state MDP assumption.

8 Revisiting the Assumptions I

Rational behavior can be defined as a Nash equilibrium of a multiple-selves game for a general POMDP with any type of discounting of future rewards. In this article, we analyzed the time-invariant Nash equilibrium solution in the limit of NOMDP with infinite discounting. In this section, we discuss how relaxing these assumptions affects choice behavior.

8.1 Infinite Temporal Discounting. Infinite discounting assumption implies that the agent is myopic and thus ignores the future consequences of his actions. In the multiple-selves framework, this assumption implies that the player ignores the effect of his chosen action on future players when forming his policy. As we demonstrated in the addiction model example, this leads to a prisoner's dilemma–like conflict between the selves, which may result in a reduced average reward to all players. If the objective function of the agent is to maximize the sum of current and all future rewards, the Nash equilibrium solution in the multiple-selves game has a simple interpretation. Consider the case of an agent that has no information about the state (NOMDP) and does not discount future rewards.[8] In this case, each self in the multiple-selves game shares his rewards with the past selves, and therefore the latter have an incentive to cooperate with him. It can be shown that in this case, $\mathbf{p}_{\max}$ is a time-invariant Nash equilibrium of the multiple-selves game (Loewenstein, Prelec, & Seung, 2008). For example, in the addiction example, this cooperation reduces the probability of "taking drugs" ($A_1 = 1$) in the Nash equilibrium because "taking drugs" reduces the expected reward of future selves. In contrast, in the two-armed bandit reward schedule in which actions have no future consequences, the Nash equilibrium solution remains unchanged.

8.2 The NOMDP Assumption. The NOMDP assumption implies that the agent does not take into account the sequence of his past actions when choosing his current action. In the multiple-selves framework, this implies that players do not take into account the actions made by past players when choosing their actions. If the NOMDP assumption is relaxed and we assume that the agent has full information about the state (making it a fully observable MDP problem), then in the Nash equilibrium solution, the action chosen by a self will, in general, depend on the actions chosen by past selves. If the objective function of the agent is to maximize the expectation value of the sum of present and exponentially discounted future rewards, the Nash equilibrium solution of this multiple-selves game can be found

[8] More precisely, we take the limit of no temporal discounting, for example, by assuming that the objective of the subject is to maximize the sum of F future rewards and take the limit of $F \to \infty$.

by solving the Bellman optimality equation (Sutton & Barto, 1998). In this solution, the chosen action of each self is, in general, a deterministic function of the actions made by past selves. It can also be shown that for the reward schedule of equation 6.1, the sequence of actions made by the different players at the Nash equilibrium is in general periodic.

9 Discussion

The unpredictability of individual choices and the matching law are empirical laws of behavior. The goal of this article was to see whether this behavior can be explained by a nomative theory that views a subject as a deliberative agent, with full information about the reward schedule but only partial information about the state of the world, that maximizes a subjective utility function. We assumed that the agent has no knowledge of the state of the world and that the agent discounts the future completely. With these assumptions, we have shown that the time-invariant Nash equilibria of the multiple-selves game are consistent with both the unpredictability of individual choices and with the matching law.

9.1 Not All Matching Behaviors Are Nash Equilibria. While the time-invariant Nash equilibria of the multiple-selves game are consistent with the matching law, the converse is not true. There are behaviors that satisfy the matching law but are not Nash equilibria, because of two important differences. First, the matching law, equation 1.1, describes the aggregate behavior and thus is indifferent to the question of how individual actions are chosen. In contrast, the Nash equilibrium solution focuses on individual actions. It predicts that individual actions are chosen as though by the tossing of a biased coin, and thus choices in consecutive trials are expected to be uncorrelated. Therefore, an experiment in which individual choices are correlated will not be consistent with the Nash equilibrium solution but can still satisfy the empirical matching law, equation 1.1. Second, as was demonstrated in the examples above, even in the aggregate sense, many of the solutions of equation 1.1 are not Nash equilibria of the game, for example, choosing the less rewarding action in a two-armed bandit experiments. Therefore, the family of the time-invariant Nash equilibria of the multiple-selves game is nested within the family of matching behaviors, the solutions of equation 1.1.

To the best of our knowledge, the experimentally observed matching behavior is restricted to the family of Nash equilibria of the multiple-selves game. First, the experimentally observed temporal correlations between actions are typically weak and behavior is well characterized by a single parameter, the probability of choice (Bush & Mosteller, 1955; Sugrue et al., 2004; Glimcher, 2003). Second, the expected reward from the nonchosen alternatives does not exceed that of the chosen alternatives. Thus, we argue

that the time-invariant Nash equilibria of the multiple-selves game is a better predictor of behavior than the empirical matching law, equation 1.1.

9.2 Revisiting the Assumptions II. In this article, we considered choice behavior of a rational agent that has full information about the transition function f and reward function r. This implies that in laboratory experiments, the subject has full information about the reward schedule, equation 6.1. However, in real life, as well as in most experimental settings, subjects do not have direct access to the transition function f and reward function r and have to infer them from their actions and the consequent rewards. However, even inferring the state space from the observations is, in general, a very difficult task. Furthermore, even if the state space is known, it is typically very large. For example, if the reward schedule is determined by equation 6.1, the number of possible states is exponential in W. Egelman et al. (1998) studied human matching behavior with $W = 40$ and $N = 2$. The number of states in this experiment was $2^{40} \approx 10^{12}$. While it is not impossible for a human subject to keep track of his past 40 choices (though it is difficult), he cannot learn the 10^{12} entries of the reward function r and the 10^{24} entries of the transition function f. This problem is often referred to as the curse of dimensionality in reinforcement learning theory (Sutton & Barto, 1998). Learning the transition and reward functions becomes even more complicated if these functions are not stationary but change with time, as is often the case in many real-life situations, as well as in experimental settings (Sugrue et al., 2004; Gallistel et al., 2001). Thus, subjects that do remember past choices may behave as though they have no knowledge of the state and find the optimal solution for the NOMDP condition in order to speed up convergence to a stationary behavioral pattern, which, albeit suboptimal, is often not far from the optimal solution for many reward schedules.

A similar problem is faced by an agent who only weakly discounts future rewards. In order for him to choose the optimal strategy, even under the NOMDP assumption, he needs to estimate the future consequences of his actions. However, he typically does not know how far reaching the consequences of his actions may be. For example, a human subject in the Egelman et al. (1998) experiments whose objective function is to maximize the overall reward in the experiment (i.e., no discounting) should consider the effect of his current action on the reward delivered up to 40 trials in the future. This problem is known as the temporal credit assignment problem in reinforcement learning theory (Sutton & Barto, 1998). Baxter and Bartlett (2001) demonstrated that the more the future is taken into account when learning the optimal stochastic strategy, the slower is the learning. Thus, subjects that do care about future rewards may behave as though they discount the future completely in order to facilitate the learning, implicitly assuming that future rewards are only weakly affected by their current actions. In the case of the two-armed bandit reward schedule, this will facilitate the

convergence to the alternative that maximizes the total expected reward. In contrast, in the addiction model reward schedule, this will lead to a stationary behavior that is very different from the one that maximizes the overall reward. Interestingly, when explicit information about the possible effect of past actions on future rewards is given to human subjects, they shift their choice preference in the direction of $\mathbf{p}_{\max}$ (Herrnstein et al., 1986), which is the Nash equilibrium of a game between multiple selves that do not discount future rewards.

9.3 NOMDP Versus One-State World. The one-state world assumption and the NOMDP assumption with infinite temporal discounting have interesting similarities. The optimal strategy in both cases ignores the future consequences of choices, in the former case because there are no consequences and in the latter because the subject is indifferent to these consequences. Moreover, in both cases, the optimal strategy is time invariant, in the former case because there is only one state and in the latter case because the state of the world is unknown. However, in general, a policy in which choices are made by tossing a biased coin with $0 < p < 1$ can be considered optimal only in the case of an NOMDP.

Appendix: Materials and Methods

Consider a game with a countably infinite number of players, indexed by the integer t. Each player chooses from the same action space. Suppose that $\{R(t)\}_{t=-\infty}^{\infty}$ are the payoffs for the pure strategy $\{\mathbf{A}(t)\}_{t=-\infty}^{\infty}$. If the game is time invariant, then $\{R(t-c)\}_{t=-\infty}^{\infty}$ are the payoffs for the pure strategy $\{\mathbf{A}(t-c)\}_{t=-\infty}^{\infty}$, for any shift c.

Theorem 2. *For a time-invariant game, there exists a Nash equilibrium in mixed strategies that is time invariant: all players choose from the same probability distribution* $\boldsymbol{p}$.

Proof. The proof is a simple extension of the one given by Cheng, Reeves, Vorobeychik, and Wellman (2004). Let $u(\mathbf{q}, \mathbf{p})$ be the payoff of a player who chooses mixed strategy $\mathbf{q}$, while all other players choose $\mathbf{p}$. The payoff is independent of t, since the game is time invariant. Let $g_k(\mathbf{p})$ be the gain (if any) of a player who chooses pure strategy $\mathbf{e}_k$ when all other players play $\mathbf{p}$:

$$g_k(\mathbf{p}) \equiv \max\left\{0, u(\mathbf{e}_k, \mathbf{p}) - u(\mathbf{p}, \mathbf{p})\right\}$$

Note that $g_k(\mathbf{p}) = 0$ for all k if and only if all players choosing $\mathbf{p}$ are a Nash equilibrium.

Define a map from the probability simplex into itself:

$$\Phi(\mathbf{p}) \equiv \frac{\mathbf{p} + \sum_k g_k(\mathbf{p})\,\mathbf{e}_k}{1 + \sum_k g_k(\mathbf{p})}. \tag{A.1}$$

(It is straightforward to verify that $\Phi(\mathbf{p})$ is a normalized probability distribution.) Since Φ is a continuous mapping of a compact convex set, it has a fixed point $\mathbf{p}^*$ by Brouwer's theorem. We now prove that $\mathbf{p}^*$ must be a Nash equilibrium. Since

$$u(\mathbf{p}^*, \mathbf{p}^*) = \sum_k p_k^* u(\mathbf{e}_k, \mathbf{p}^*),$$

there is some pure strategy $\mathbf{e}_w$ for which $p_w^* > 0$ and $u(\mathbf{e}_w, \mathbf{p}^*) \leq u(\mathbf{p}^*, \mathbf{p}^*)$. For this w, equation A.1 implies that

$$p_w^* = \frac{p_w^* + g_w(\mathbf{p}^*)}{1 + \sum_k g_k(\mathbf{p}^*)} = \frac{p_w^*}{1 + \sum_k g_k(\mathbf{p}^*)}.$$

This implies $\sum_k g_k(\mathbf{p}^*) = 0$ or, equivalently, $g_k(\mathbf{p}^*) = 0$ for all k. Therefore all players choosing $\mathbf{p}^*$ are a Nash equilibrium.

Glossary

Normal form game. A game in normal form has three elements:

1. A set of players. In the multiple-selves game, there is an infinite number of players denoted by t, $t \in (-\infty, \infty)$.
2. A pure strategy space for every player. In the multiple-selves game, there are N pure strategies for every player, corresponding to the N actions available to the subject.
3. A payoff function for every player that specifies his payoff for every possible combination of pure strategies employed by all players. In the multiple-selves game with infinite temporal discounting, the payoff function of a player depends on his strategy, as well as the strategies preceding players, such that the payoff to player t is equal to $r(\mathbf{A}(t), \mathbf{A}(t-1), \mathbf{A}(t-2), \ldots)$, where $\mathbf{A}(t)$ is the pure strategy employed by player t.

Mixed strategy. A mixed strategy is a probability distribution over pure strategies. Each player's randomization is statistically independent of those of other players, and the payoffs to a profile of mixed strategies are the expected values of the corresponding pure strategy payoffs.

Nash equilibrium. A Nash equilibrium is a profile of strategies such that no player can increase his payoff by changing only his own strategy. In this sense, each player's strategy is an optimal response to the other players' strategies (Fudenberg & Tirole, 1991).

Acknowledgments

Y.L. was supported by the Israel Science Foundation (grant no. 868/08); H.S.S was supported by the Howard Hughes Medical Institute.

References

Ainslie, G. (1992). *Picoeconomics: The strategic interaction of successive motivational states within the person*. Cambridge: Cambridge University Press.

Aumann, R. J., Hart, S., & Perry, M. (1997). The absent-minded driver. *Games and Economic Behavior, 20*(1), 102–116.

Baxter, J., & Bartlett, P. L. (2001). Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research, 15*(4), 319–350.

Bush, R. R., & Mosteller, F. (1955). *Stochastic models for learning*. Hoboken, NJ: Wiley.

Cheng, S. F., Reeves, D. M., Vorobeychik, Y., & Wellman, M. P. (2004). Notes on equilibria in symmetric games. In *Proceedings of the AAMAS-04 Workshop on Game-Theoretic and Decision-Theoretic Agents*. New York: ACM Press.

Davison, M., & McCarthy, D. (1988). *The matching law: A research review*. Mahwah, NJ: Erlbaum.

Egelman, D. M., Person, C., & Montague, P. R. (1998). A computational role for dopamine delivery in human decision-making. *Journal of Cognitive Neuroscience, 10*(5), 623–630.

Frederick, S., Loewenstein, G., & ODonoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature, 40*(2), 351–401.

Fudenberg, D., & Levine, D. K. (1998). *The theory of learning in games*. Cambridge, MA: MIT Press.

Fudenberg, D., & Tirole, J. (1991). *Game theory*. Cambridge, MA: MIT Press.

Gallistel, C. R. (1990). *The organization of learning*. Cambridge, MA: MIT Press.

Gallistel, C. R., Mark, T. A., King, A., & Latham, P. E. (2001). The rat approximates an ideal detector of changes in rates of reward: Implications for the law of effect. *Journal of Experimental Psychology: Animal Behavior Processes, 27*(4), 354–372.

Gallistel, C. R., Mark, T. A., King, A., & Latham, P. (2002). A test of Gibbon's feed-forward model of matching. *Learning and Motivation, 33*(1), 46–62.

Gibbon, J. (1995). Dynamics of time matching: Arousal makes better seem worse. *Psychonomic Bulletin and Review, 2*(2), 208–215.

Gilboa, I. (1997). A comment on the absent-minded driver paradox. *Games and Economic Behavior, 20*(1), 25–30.

Glimcher, P. W. (2003). *Decisions, uncertainty, and the brain: The science of neuroeconomics*. Cambridge, MA: MIT Press.

Hammond, P. J. (1997). Rationality in economics. *Rivista internazionale di scienze sociali*, *105*(3), 247–288.

Hendon, E., Jacobsen, H. J., & Sloth, B. (1996). The one-shot-deviation principle for sequential rationality. *Games and Economic Behavior, 12*(2), 274–282.

Herrnstein, R. J., Loewenstein, G. F., Prelec, D., & Vaughan Jr., W. (1993). Utility maximization and melioration: Internalities in individual choice. *Journal of Behavioral Decision Making, 6*(3), 149–185.

Herrnstein, R. J., & Loveland, D. H. (2001). Maximizing and matching on concurrent ratio schedules. *J. Exp. Anal. Behav., 1975*, 107–116.

Herrnstein, R. J., & Prelec, D. (1991). Melioration: A theory of distributed choice. *Journal of Economic Perspectives, 5*(3), 137–156.

Herrnstein, R. J., Prelec, D., & Vaughan, Jr., W. (1986). An intra-personal prisoners' dilemma. In *IX Symposium on the Quantitative Analysis of Behavior.* N.p.

Heyman, G. M., & Dunn, B. (2002). Decision biases and persistent illicit drug use: An experimental study of distributed choice and addiction. *Drug and Alcohol Dependence, 67*(2), 193–203.

Heyman, G. M., & Luce, R. D. (1979). Operant matching is not a logical consequence of maximizing reinforcement rate. *Animal Learning and Behavior, 7*, 133–140.

Hofbauer, J., & Sigmund, K. (2003). Evolutionary game dynamics. *American Mathematical Society, 40*(4), 479–519.

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics, 112*(2), 443–477.

Loewenstein, Y. (2008). Robustness of learning that is based on covariance-driven synaptic plasticity. *PLoS Computational Biology, 4*(3), e1000007.

Loewenstein, Y., Prelec, D., & Seung, H. S. (2007). A game theoretical approach to the matching law: Operant matching is a Nash equilibrium of an interpersonal game. *Abstract viewer Itinerary planner.* Washington, DC: Society for Neuroscience.

Loewenstein, Y., Prelec, D., & Seung, H. S. (2008, December 1). *Dynamic utility maximization by reinforcement thinking*. Unpublished manuscript.

Loewenstein, Y., & Seung, H. S. (2006). Operant matching is a generic outcome of synaptic plasticity based on the covariance between reward and neural activity. *Proceedings of the National Academy of Sciences, 103*(41), 15224–15229.

Mazur, J. E. (1981). Optimization theory fails to predict performance of pigeons in a two-response situation. *Science, 214*(4522), 823–825.

McAllester, D., & Singh, S. (1999). Approximate planning for factored POMDPs using belief state simplification. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (pp. 409–416). N.p.: AUAI Press

Neiman, T., & Loewenstein, Y. (2007). A dynamic model for matching behavior that is based on the covariance of reward and action. *Neural Plasticity, 2007*, 79.

Piccione, M., & Rubinstein, A. (1997). On the interpretation of decision problems with imperfect recall. *Games and Economic Behavior, 20*(1), 3–24.

Rachlin, H., & Laibson, D. I. (Eds.). (1997): *The matching law: Papers in psychology and economics*. Cambridge, MA: Harvard University Press.

Sakai, Y., & Fukai, T. (2008). The actor-critic learning is behind the matching law: Matching versus optimal behaviors. *Neural Computation, 20*(1), 227–251.

Seung, H. S. (2003). Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. *Neuron, 40*(6), 1063–1073.

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making, 15*(3), 233–250.

Soltani, A., & Wang, X. J. (2006). A biophysically based neural model of matching law behavior: Melioration by stochastic synapses. *Journal of Neuroscience, 26*(14), 3731–3744.

Strotz, R. H. (1955). Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies, 23*(3), 165–180.

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2004). Matching behavior and the representation of value in the parietal cortex. *Science, 304*(5678), 1782–1787.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Vaughan, W. (1981). Melioration, matching, and maximization. *J. Exp. Anal. Behav., 36*(2), 141–149.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys, 14*(1), 101–118.

Vyse, S. A., & Belke, T. W. (1992). Maximizing versus matching on concurrent variable-interval schedules. *J. Exp. Anal. Behav., 58*(2), 325–334.

Received September 3, 2008; accepted February 13, 2009.